Data Driven Reading Comprehension

Phil Blunsom

In collaboration with Karl Moritz Hermann, Tomáš Kočiský, Ed Grefenstette and the DeepMind Natural Language Group
The DeepMind Language Group
We aim to build models that can read a text, represent the information contained within it, and answer questions based on this representation.

There are two broad motivations for doing this,

1. To build QA applications or products,
2. To evaluate language understanding algorithms.
James the Turtle was always getting in trouble. Sometimes he’d reach into the freezer and empty out all the food. Other times he’d sled on the deck and get a splinter. His aunt Jane tried as hard as she could to keep him out of trouble, but he was sneaky and got into lots of trouble behind her back. One day, James thought he would go into town and see what kind of trouble he could get into. He went to the grocery store and pulled all the pudding off the shelves and ate two jars. Then he walked to the fast food restaurant and ordered 15 bags of fries. He didn’t pay, and instead headed home. …

Where did James go after he went to the grocery store?

1. his deck
2. his freezer
3. a fast food restaurant
4. his room
The CNN and Daily Mail datasets: aims

The CNN and Daily Mail websites provide paraphrase summary sentences for each full news story.

Hundreds of thousands of documents
Millions of context-query pairs
Hundreds of entities

1Hermann et al. Teaching machines to read and comprehend. NIPS 2015
The CNN and Daily Mail datasets: large scale RC

MC Test

500 stories, 2k questions.

CNN and Daily Mail Corpora

~300k stories and >1M questions.
The BBC producer allegedly struck by Jeremy Clarkson will not press charges against the “Top Gear” host, his lawyer said Friday. Clarkson, who hosted one of the most-watched television shows in the world, was dropped by the BBC Wednesday after an internal investigation by the British broadcaster found he had subjected producer Oisin Tymon “to an unprovoked physical and verbal attack.” …

**Cloze-style question:**

**Query:** Producer X will not press charges against Jeremy Clarkson, his lawyer says.

**Answer:** Oisin Tymon
The CNN and Daily Mail datasets: the data

From the Daily Mail:

- The hi-tech bra that helps you beat breast X
- Could Saccharin help beat X?
- Can fish oils help fight prostate X?

Any n-gram language model trained on the Daily Mail would correctly predict (X = cancer)
The CNN and Daily Mail datasets: anonymisation

We aimed to design the task to avoid shortcuts such as QA by language modelling or correlation:

*(CNN)* New Zealand are on course for a first ever World Cup title after a thrilling semifinal victory over South Africa, secured off the penultimate ball of the match.

Chasing an adjusted target of 298 in just 43 overs after a rain interrupted the match at Eden Park, Grant Elliott hit a six right at the death to confirm victory and send the Auckland crowd into raptures. It is the first time they have ever reached a world cup final.

**Question:** ____ reach cricket World Cup final?

**Answer:** New Zealand

*(ent23)* ent7 are on course for a first ever ent15 title after a thrilling semifinal victory over ent34, secured off the penultimate ball of the match.

Chasing an adjusted target of 298 in just 43 overs after a rain interrupted the match at ent12, ent17 hit a six right at the death to confirm victory and send the ent83 crowd into raptures. It is the first time they have ever reached a ent15 final.

**Question:** ____ reach ent3 ent15 final?

**Answer:** ent7
The CNN and Daily Mail datasets: models

We proposed a simple attention based approach:

- Separate encodings for query and context tokens
- Attend over context token encodings
- Predict based on joint weighted attention and query representation
The CNN and Daily Mail datasets: the good and bad

Good: we recognised that there must be a level of indirection between annotators producing questions and the text from which the questions are answered.

Bad: many of the automatically generated questions are of poor quality or ambiguous.¹

¹Chen et al. A Thorough Examination of the CNN/Daily Mail Reading Comprehension Task. ACL 2016
The CNN and Daily Mail datasets: the good and bad

Good: we aimed to factor out world knowledge through entity anonymisation so models could not rely on correlations rather than understanding.

Bad: The generation process and entity anonymisation reduced the task to multiple choice and introduced additional noise.
The CNN and Daily Mail datasets: the good and bad

Good: posing reading comprehension as a large scale conditional modelling task made it accessible to machine learning researchers, generating a great deal of subsequent research.

Bad: while this approach is reasonable for building applications, it is entirely the wrong way to develop and evaluate natural language understanding.
Desiderata for Reading Comprehension Data sets

Applications

If our aim is to build a product or application, we must acquire data as close to the real use case as possible, i.e. representative questions and document contexts.

If we artificially generate data we risk introducing spurious correlations, which overparameterised neural networks are excellent at exploiting.
Desiderata for Reading Comprehension Data sets

Language Understanding

Any data annotation process will introduce spurious artifacts into the data.

Performance on a language understanding evaluation can thus be factored into two components, 1) that which measures true understanding, 2) and that which captures overfitting to the artifacts.

If our aim is to evaluate language understanding systems we must not train on data collected with the same annotation process as our evaluation set.
In the 1960s, a series of discoveries, the most important of which was seafloor spreading, showed that the Earth's lithosphere, which includes the crust and rigid uppermost portion of the upper mantle, is separated into a number of tectonic plates that move across the plastically deforming, solid, upper mantle, which is called the asthenosphere. There is an intimate coupling between the movement of the plates on the surface and the convection of...

Question:
Which parts of the Earth are included in the lithosphere?

1Rajpurkar et al. SQuAD: 100,000+ Questions for Machine Comprehension of Text. EMNLP 2016
Stanford Question Answering Dataset (SQuAD)

Good: Very scalable annotation process that can cheaply generate large numbers of questions per article.

Bad: Annotating questions directly from the context passages strongly skews the data distribution. The task then becomes reverse engineering the annotators, rather than language understanding.

1https://rajpurkar.github.io/SQuAD-explorer/
Stanford Question Answering Dataset (SQuAD)

Good: The online leaderboard allows easy benchmarking of systems and motivates competition.

Bad: Answers as spans reduces the task to multiple choice, and doesn’t allow questions with answers latent in the text.

1https://rajpurkar.github.io/SQuAD-explorer/
Stanford Question Answering Dataset (SQuAD)

SQUAD provides a great resource for experimenting with machine learning models.

However, just like the CNN/DailyMail corpus, it does not satisfy the requirements for building applications, nor for evaluating language understanding systems.

1https://rajpurkar.github.io/SQuAD-explorer/
Questions are mined from a search engine and matched with candidate answer passages using IR techniques.

Answers are not restricted to be subspans of the documents, and some questions are not answerable from the context.
MS Marco

Good: The reliance on real queries creates a much more useful resource for those interested in applications.

Bad: People rarely ask interesting questions of search engines, and the use of IR techniques to collect candidate passages limits the usefulness of this dataset for evaluating language understanding.

Good: Unrestricted answers allow a greater range of questions.

Bad: How to evaluate freeform answers is an unsolved problem. Bleu is not the answer!

Narrative QA: aims

Understanding language goes beyond reading and answering literal questions on factual content.

Narratives present many interesting challenges, requiring models to represent and reason over characters and temporal relationships.

1Kocisky et al. The NarrativeQA Reading Comprehension Challenge. TACL 2018
Narrative QA: construction

Documents are **books** and **movie scripts**
- Complex, long, self-contained narratives
- Contain dialogue

Questions from **abstractive summaries**
- Summary → 30 questions with 2 answers each

Answers are human generated

---

1Kocisky et al. The NarrativeQA Reading Comprehension Challenge. TACL 2018
In what year did Rogers awaken from his deep slumber? 

**Answer:** 2419

...Rogers remained in sleep for 492 years. **He awakes in 2419** and...

I should state therefore, that I, Anthony Rogers, am, so far as I know, the only man alive whose normal span of eighty-one years of life has been spread over a period of 573 years. To be precise, **I lived the first twenty-nine years of my life between 1898 and 1927; the other fifty-two since 2419.** The gap between these two, a period of nearly five hundred years, I spent in a state of suspended animation, free from the ravages of katabolic processes, and without any apparent effect on my physical or mental faculties. When I began my long sleep, man had just begun his real conquest of the air...
Question: How is Oscar related to Dana?
Answer: He is her son

Summary snippet: ...Peter's former girlfriend Dana Barrett has had a son, Oscar...

Story snippet:
DANA (setting the wheel brakes on the buggy) Thank you, Frank. I'll get the hang of this eventually.

She continues digging in her purse while Frank leans over the buggy and makes funny faces at the baby, OSCAR, a very cute nine-month old boy.

FRANK (to the baby) Hiya, Oscar. What do you say, slugger?
FRANK (to Dana) That's a good-looking kid you got there, Ms. Barrett.
## Narrative QA: composition

<table>
<thead>
<tr>
<th></th>
<th>Train</th>
<th>Validation</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td># documents</td>
<td>1,102</td>
<td>115</td>
<td>355</td>
</tr>
<tr>
<td># questions</td>
<td>32,747</td>
<td>3,461</td>
<td>10,557</td>
</tr>
<tr>
<td>Avg. # tokens in summaries</td>
<td>659</td>
<td>638</td>
<td>654</td>
</tr>
<tr>
<td>Avg. # tokens in scripts</td>
<td>29,934</td>
<td>29,515</td>
<td>29,900</td>
</tr>
<tr>
<td>Avg. # tokens in books</td>
<td>95,632</td>
<td>94,506</td>
<td>86,329</td>
</tr>
<tr>
<td>Max # tokens</td>
<td>430,061</td>
<td>418,265</td>
<td>404,641</td>
</tr>
</tbody>
</table>
Narrative QA: question categories

Why doesn’t Pozdnyshev run after the violinist? [Why/reason 9.4%]
How did Jake survive being shot? [How/method 8.1%]
When does Reiko realize the curse is still unbroken? [Event 4.4%]
What is Nora’s relationship to Michael? [Relation 1.3%]

Person 30.5%, Description 24.5%, Location 9.7%, Entity 4.0%, Object 3.7% Numeric 3.0%, Duration 1.7%
Narrative QA: benchmarks

We evaluated models with Bleu, Rouge, and Mean Reciprocal Rank. None of these are ideal.

Models that have performed well on tasks such as MS Marco also give similar performance when answering questions directly from the summaries.

Such models do not scale to the task of answering questions from full narratives, so we experimented with an initial IR step to retrieve candidate passages.

All the models we tried were unable to answer a significant number of questions posed on the full narratives.
Narrative QA: the good and bad

Good: A challenging evaluation that tests a range of language understanding, particularly temporal aspects of narrative, and also scalability as current models cannot represent and reason over full narratives.

Bad: Performing well on this task is clearly well beyond current models, both representationally and computationally. As such it will be hard for researchers to hill climb on this evaluation.
Narrative QA: the good and bad

**Good:** The relatively small number of narratives for training models forces researchers to approach this task from a transfer learning perspective.

**Bad:** The relatively small number of narratives means that this dataset is not of immediate use for those wanting to build supervised models for applications.
Transfer learning and Sample Complexity

We need to move away from training narrow supervised RC language understanding models towards a transfer learning paradigm.

This will require to us to let go of our CL obsession with single point accuracy comparisons and embrace sample complexity as the true goal.
Transfer learning and Sample Complexity

Multitask learning is not the answer.

It is easy to grow neural networks with the number of tasks, thus allowing models to overtrain to the idiosyncrasies of each task.
Summary

The marriage of machine learning models, large data sets, and reading comprehension has produced a vibrant research environment.

We must move beyond this paradigm, embracing a variety of language and focusing on representation and inference over local context and correlation.
Thank You.

Data Driven Reading Comprehension

Phil Blunsom

In collaboration with Karl Moritz Hermann, Tomáš Kočiský, Ed Grefenstette and the DeepMind Natural Language Group